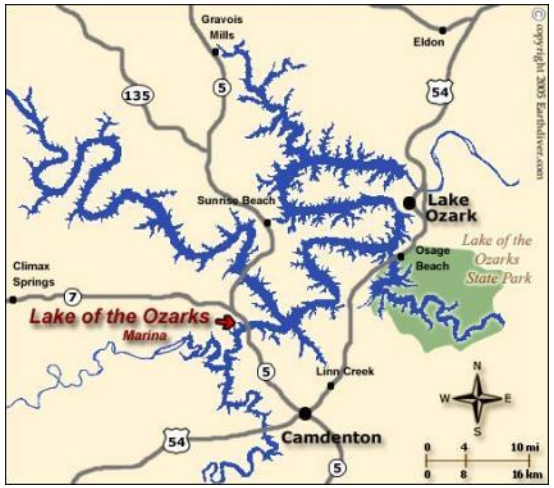
Bridging the Gaps: Satellites to Science and Desktop to the Cloud

Catharine van Ingen
Partner Architect
eScience Group, Microsoft Research

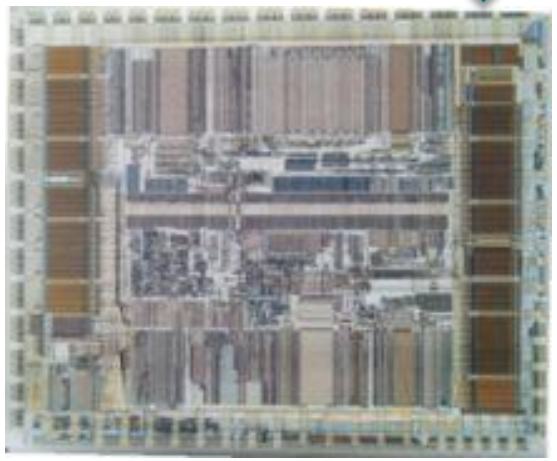
I've come full circle



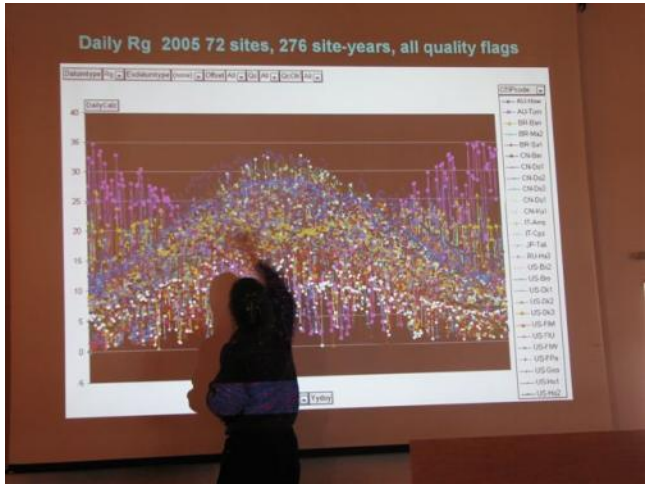
Modeler

Experimentalist

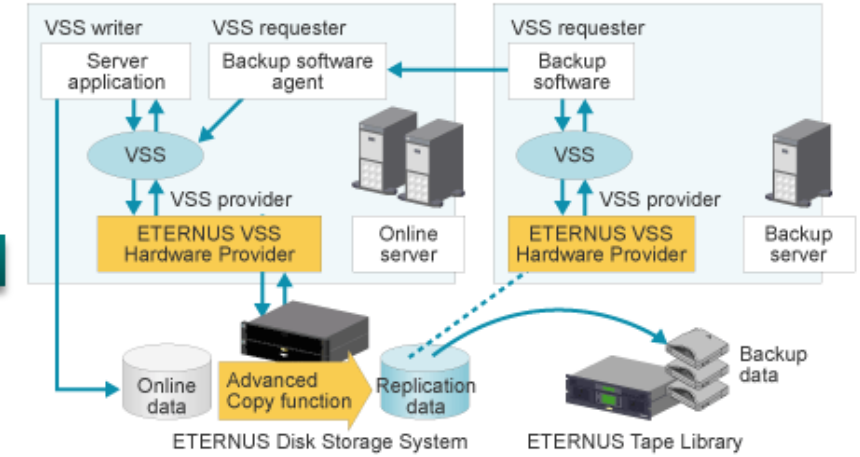
Detector integrator



Computer architect



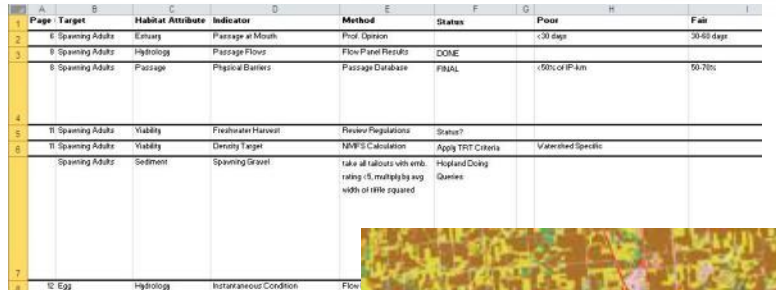
Environmental eScientist

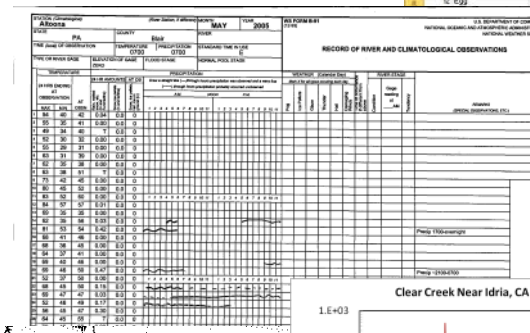


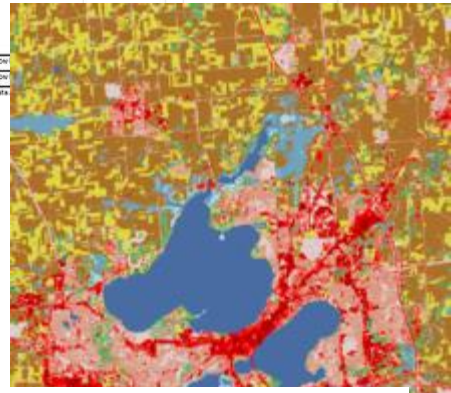
Storage architect

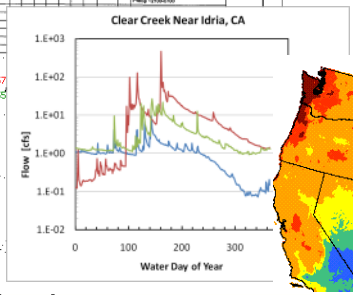
The Data is Out There

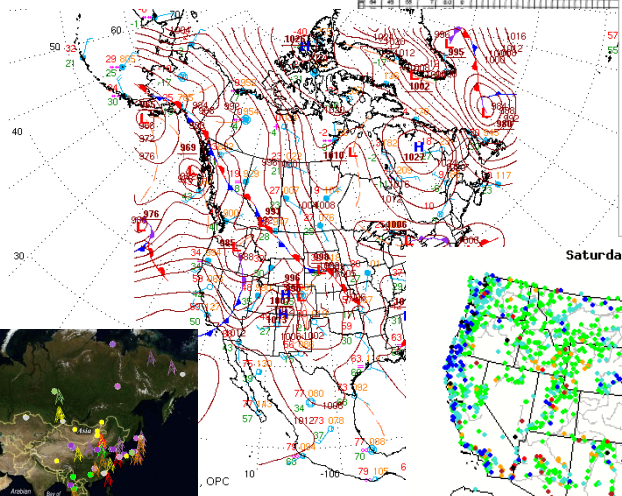
- National and International Datasets
  - USGS National Water Information System
  - NOAA National Climatic Data Center
  - FLUXNET Network
  - Satellite data (e.g. MODIS)
- Local Datasets
  - Local Agencies
  - Companies (e.g. Timber)
  - Ecology Organizations
  - Individual Researchers

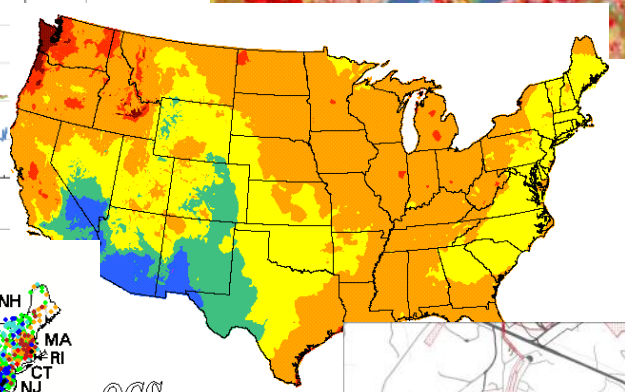


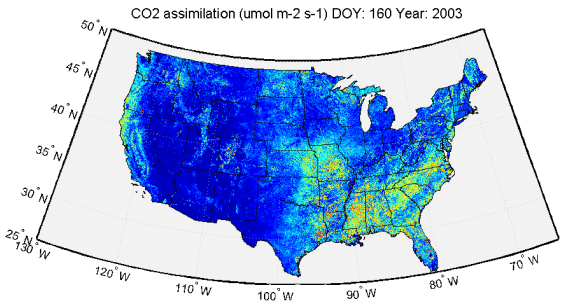


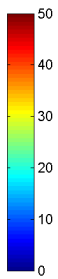


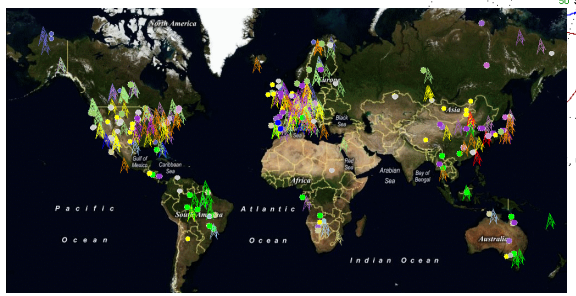


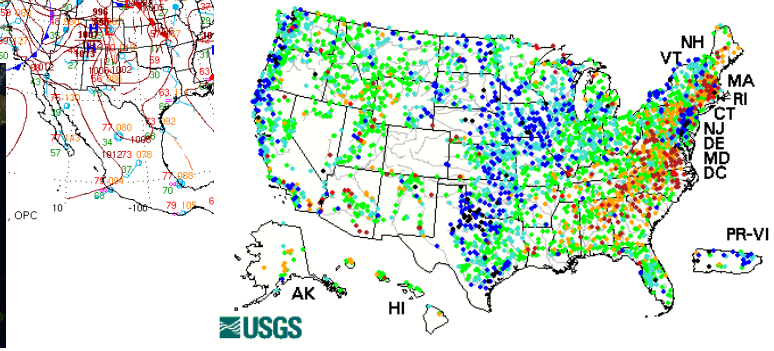










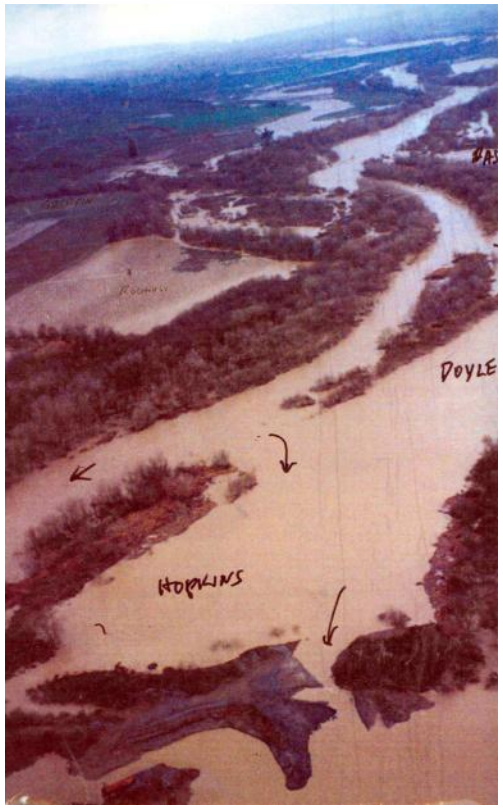


Data Variety – The Spice of Life

Manual Measurement

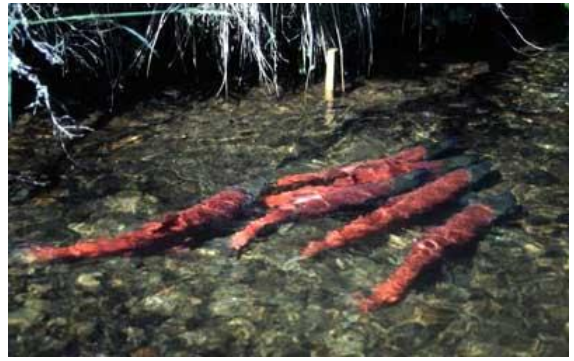
Automated Measurement

Sample Collection

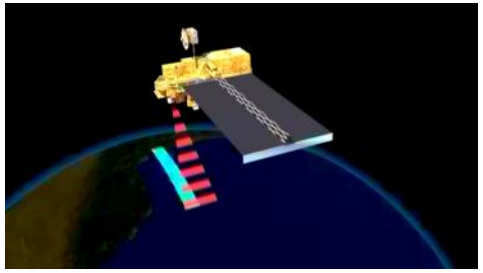


Historical Photographs

Typing

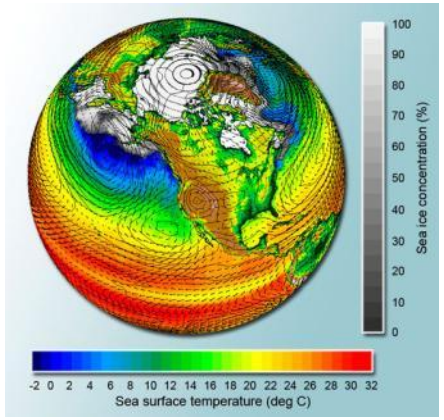


Counting



Satellite

Aircraft Surveys



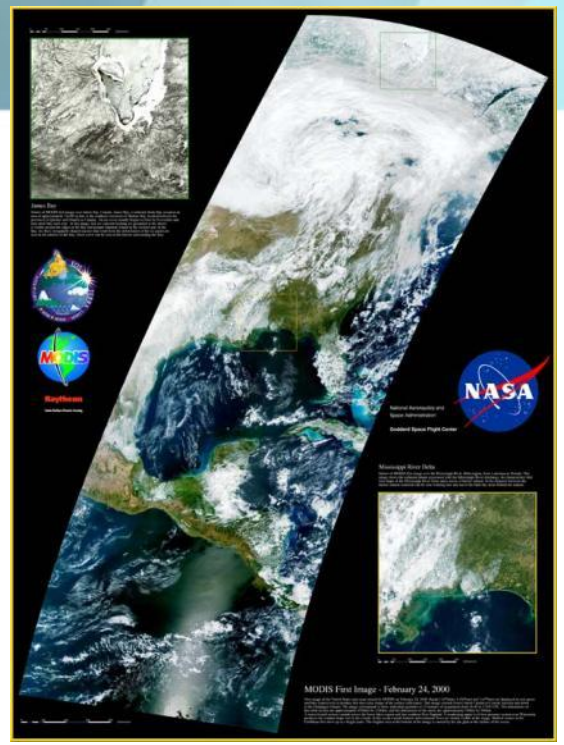
Model Output

Relatively Ubiquitous Motes

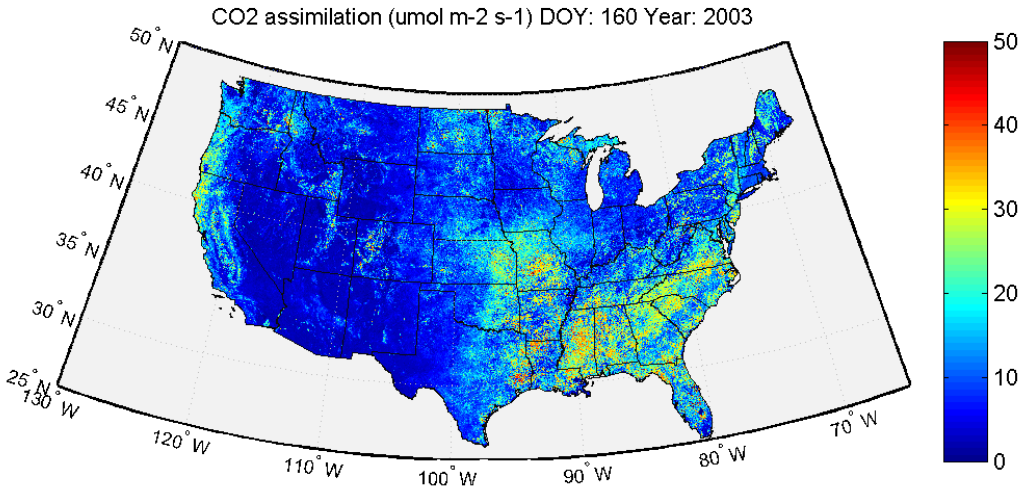
Toward a Scalable Environmental Data Practice

I resolve to stop accumulating and being the infinitely more serious and difficult task of wise distribution.
Andrew Carnegie

Environmental Remote Sensing Data



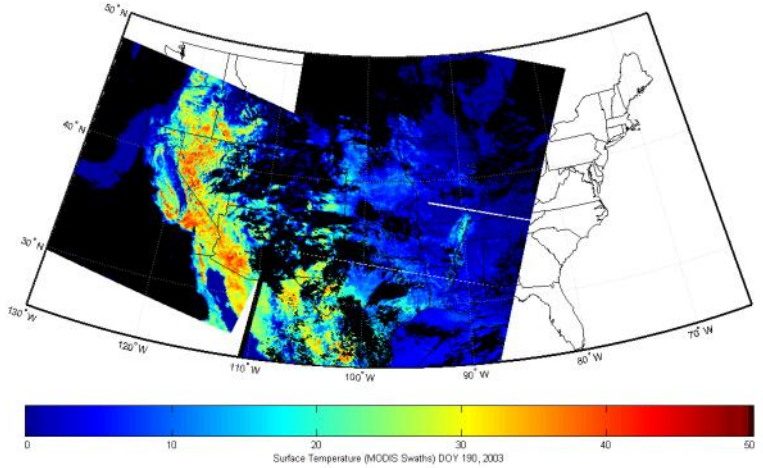
- Time series raster data
  - Over some period of time at some time frequency at some spatial granularity over some spatial area
  - Conversion from L0 data to L2 and beyond as well as reprojections still require specialized skills
  - Similar, but dirtier, than model output
- Can be “cut out” to create virtual sensors
- Today: PBs (L0) to TBs (L2+)



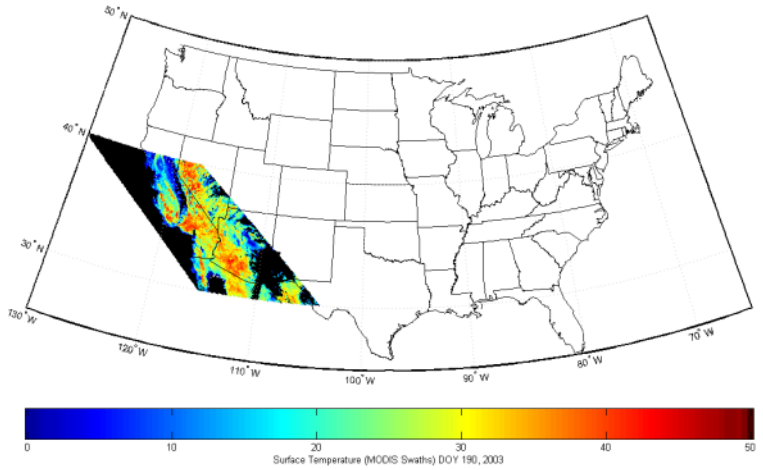
Challenges: size and data formats

Tiling : Do Scientists Have to be Computer Scientists?

- Reprojection
  - Converts one geo-spatial representation to another.
  - Examples are converting from latitude-longitude swaths to sinusoidal cells or sinusoidal cells to land-based Albers (USGS)
- Spatial resampling
  - Converts one spatial resolution to another.
  - Example is converting from 1 KM to 5 KM pixels.
- Temporal resampling
  - Converts one temporal resolution to another.
  - Example is converting from daily observation to 8 day averages.
- Gap filling
  - Assigns values to pixels without data either due to inherent data issues such as clouds, day:night observations, or missing pixels introduced by one of the above.
- Masking
  - Eliminates uninteresting or unneeded pixels.
  - Examples are eliminating pixels over the ocean when computing a land product or eliminating pixels outside a spatial feature such as a watershed.



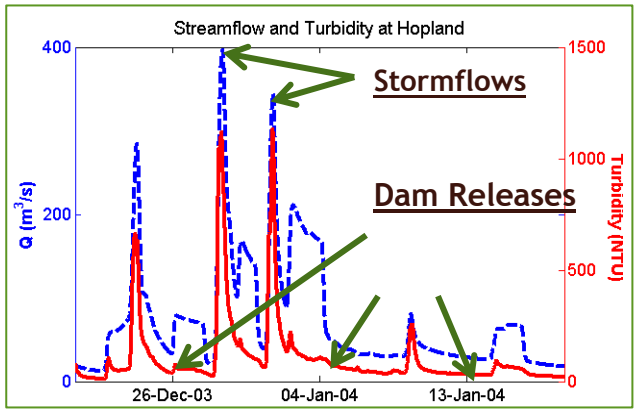
Source Data (Swath format)

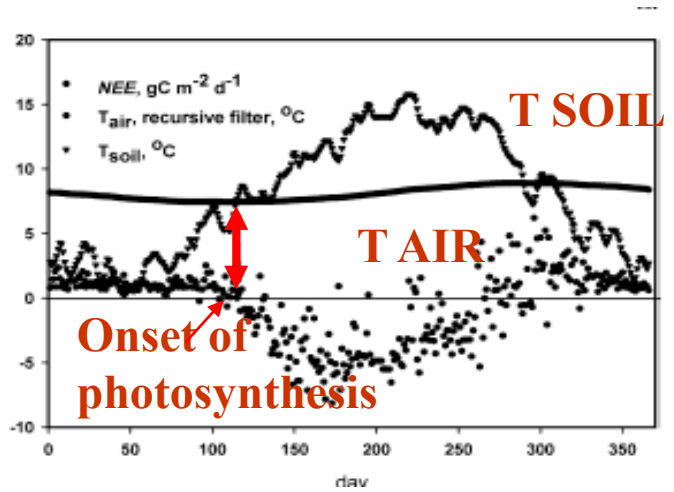


Reprojected Data (Sinusoidal format)

Environmental Sensor Data

- Time series data
  - Over some period of time at some time frequency at some spatial location.
  - May be actual measurement (L0) or derived quantities (L1+)
- (Re)calibrations, gaps and errors are a way of life.
  - Birds poop, batteries die, sensors fail.
  - Quality assessment and signal correction varies
  - Gap filling algorithms key as regular time series enable more analyzes
- Today: GBs to TBs

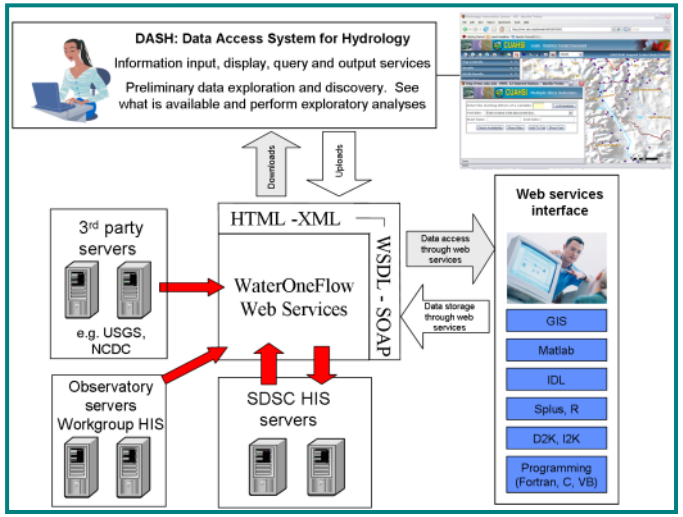


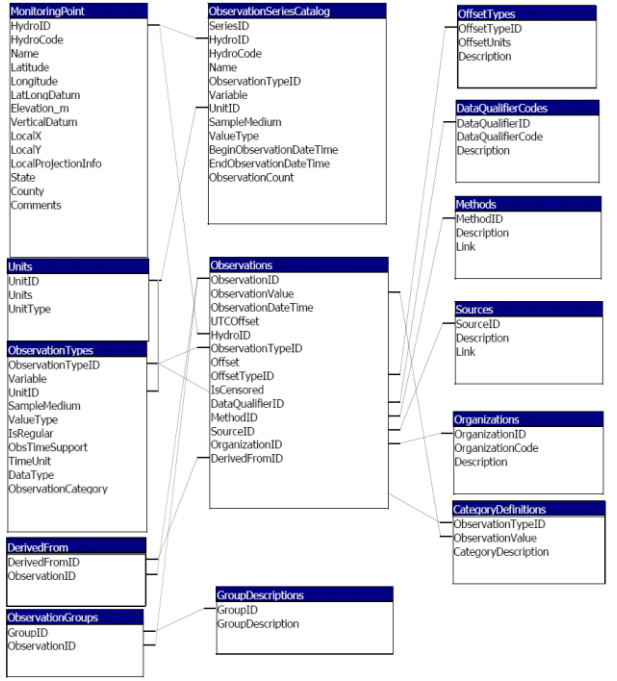


Challenges: science conversion and browsing

Sensor Databases and Web Services

- Emerging trend is that groups use databases and web services to access, curate, and republish sensor data
- Most use a mostly normalized schema with the data in the center, but moving to putting the series catalog in the center
- Example is CUAHSI ODM
  - Initially to address internet access of US agency data – too hard to find, too hard to download all the data, too hard to get "just the new data"
  - Included water quality bottle samples, a notion of data revisions
  - 11 initial research sites growing over time





http://bwc.berkeley.edu http://www.cuahsi.org

CUAHSI
universities allied for water research

Environmental Ancillary Data (Long Tail Data)

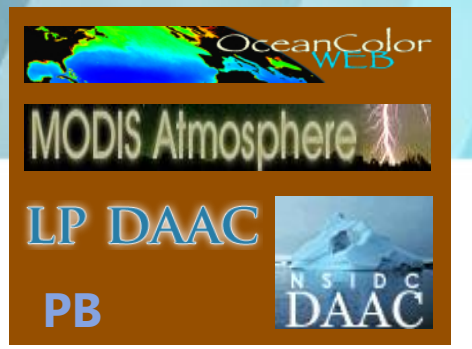
- Almost everything else!
  - ‘Constants’ such as latitude or longitude
  - Intermittent measurements such as grain size distributions or fish counts
  - Anecdotal descriptions such as “ripple” or “shaded”
  - Events such as algal blooms or leaf fall including those derived from sensor data such as “flood”
  - Disturbances such as a fire, harvest, landslide
- Not metadata such as instrument type, derivation algorithm, etc.
- Today: KBs to maybe GBs.

Challenge: harmonization and provenance

Ancillary Data is Different !

- Very hard won
  - Dig a pit or shoot an air rifle to get samples
  - Lab costs can be considerable
  - Gleaning from literature (and cross checking!)
- Very hard to curate
  - FLUXNET collection is currently ~30K numbers.
  - Often passed around in email and cut/pasted from web sites
- Very different usage patterns
  - Constant location attributes or aliases
  - Time series via splines or step functions
  - Filters for sensor data: periods before or after, sites with summer LAI > x, etc
  - Time benders: “since <event>”
- Often requires science judgment
  - Different scientists don’t always agree
  - Anecdotal reporting difficult to interpret
  - Citizen science contributions give important coverage but at quality?

Why Make These Distinctions?



- Provenance and trust widely varies
  - Data acquisition, early processing, and reporting ranges from a large government agency to individual scientists.
  - Smaller data often passed around in email; big data downloads can take days (if at all)
- Data sharing concerns and patterns vary
  - Open access followed by (non-repeatable and tedious) pre-processing
  - True science ready data set but concerns about misuse, misunderstanding particularly for hard won data.
- Computational tools differ.
  - Not everyone can get an account at a supercomputer center
  - Very large computations require engineering (error handling)
  - Space and time aren't always simple dimensions

Complex shared detector

Simple instrument (if any)

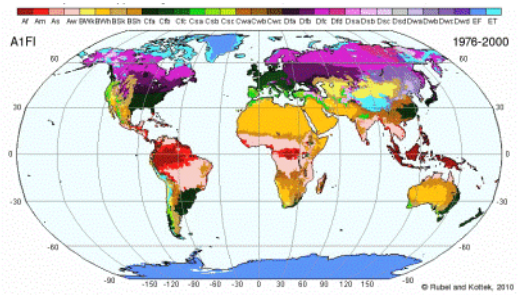
Science happens when PBs, TBs, GBs, and KBs can be mashed up simply

Complex and Heavy process by experts

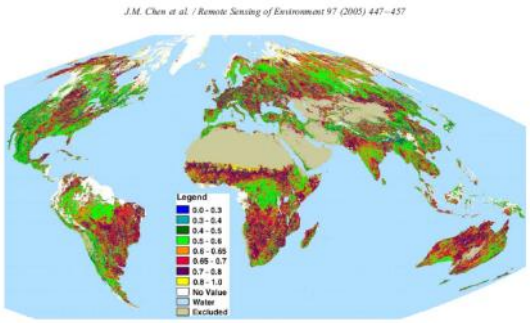
Ad hoc observations and models

Synthesizing Imagery, Sensors, Models and Field Data

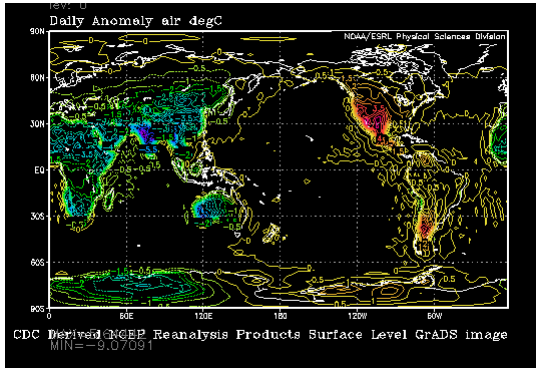
Climate classification ~1MB (1file)

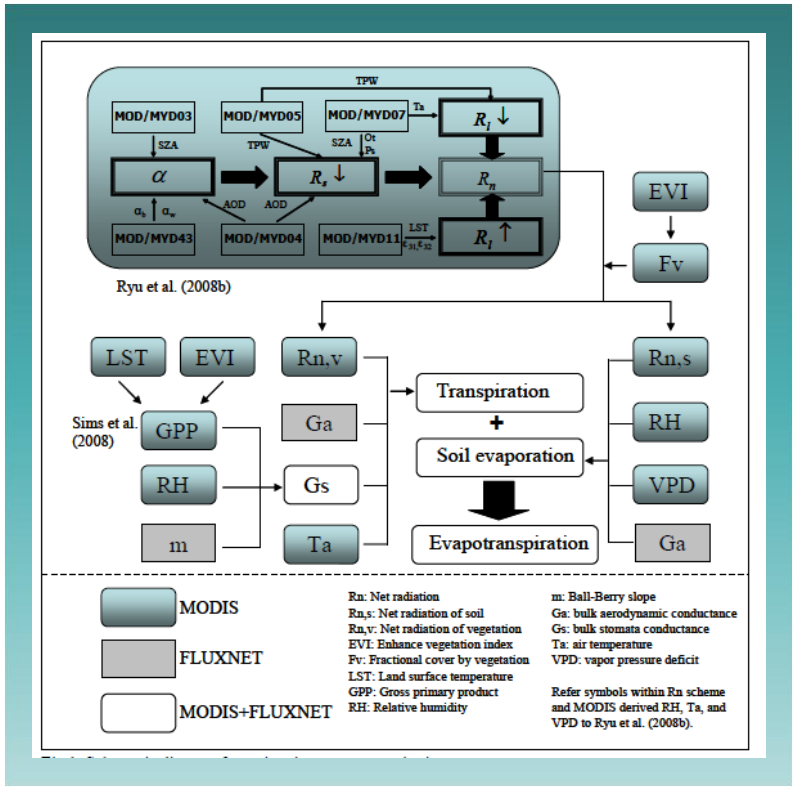


Vegetative clumping ~5MB (1file)

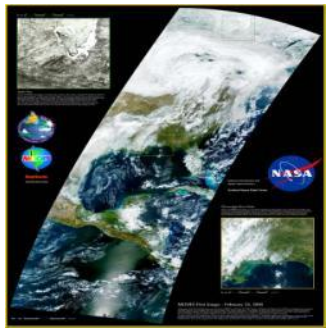


NCEP/NCAR ~100MB (4K files)





NASA MODIS imagery source archives 5 TB (600K files)



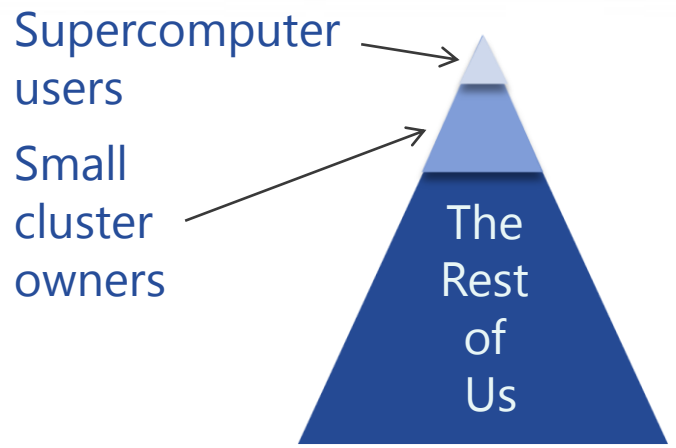
FLUXNET curated sensor dataset (30GB, 960 files)

FLUXNET curated field dataset 2 KB (1 file)

To The Cloud: Our Great Adventure

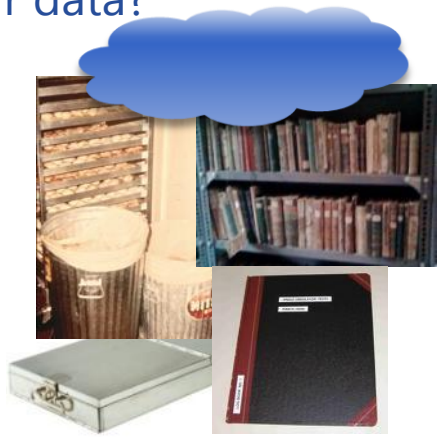
The object of all the former voyages to the South Seas undertaken by the command of his present majesty, has been the advancement of science and the increase of knowledge.
William Bligh

Bridging the Gap with the Cloud



- Barriers to Science:
  - *Resource*: compute, storage, networking, visualization capability
  - *Complexity*: specific cross-domain knowledge
  - *Tedium:* repetitive data gathering or preprocessing tasks
- With cloud computing, we can:
  - marshal needed storage and compute resources on demand without caring or knowing how that happens
  - access living curated datasets without having to find, educate, and reward a private data curator
  - run key common algorithms as Software as a Service without having to know the coding details or installing software
  - grow a given collaboration or share data and algorithms across science collaborations elastically

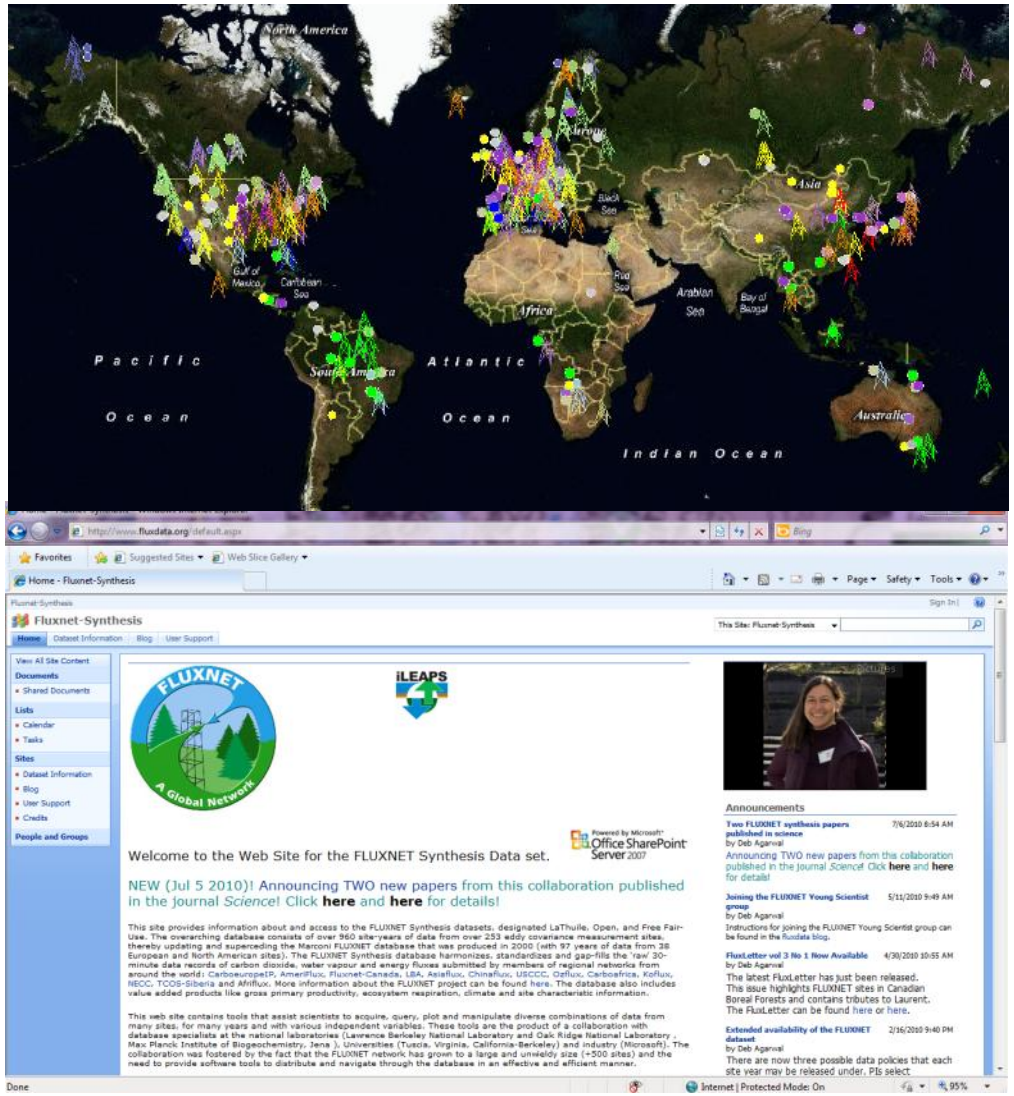
Where do you want your data?



Democratizing science analysis by fostering sharing and reuse

Cloud Classic : Fluxdata.org Learnings

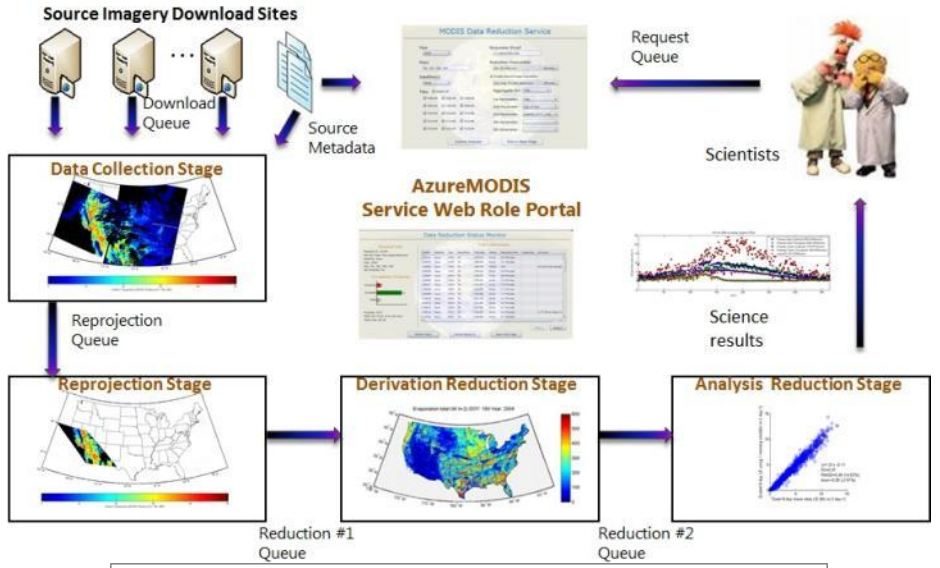
- FLUXNET is not one collaboration but rather several different sub-groups with different cultures
  - Ontologies, self-coordination, provenance, curation all mean different things to different sub-groups
  - “Barcalounger users” expect clean science ready data at all times
- Plan for data set evolution
  - Algorithms improve
  - Data can only be cleaned when used
  - Campaigns driven by specific papers or analyses give the best motivation
- Lowering the grunge barrier enables science
  - So many analyses, so little time
  - Enabling browsing for data availability enables new analyses
- Tracking publications and data usage together enables a virtuous cycle
  - Data represents both a cost and asset to the originator
  - Publication of data is more costly than access

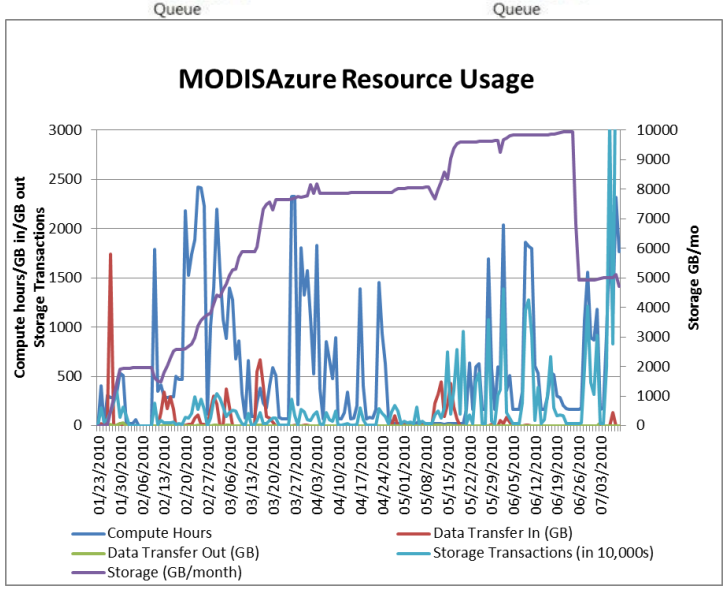


http://www.fluxdata.org

Cloud Current: MODISAzure Learnings

- Lowering the barriers to use remote sensing data can enable science
  - NASA makes the data accessible, not science ready
  - At AGU 2009, we learned that a cloud service that just made on-demand jpg mosaics would help tremendously
- On demand resources are a good match to changes in science needs
  - Computation changed over time due to scaling up (continental to global) and algorithmic sophistication
- Science and algorithm debugging benefit from the same infrastructure as both need to scale up and down
  - Debugging an algorithm on the desktop isn't enough – you have to debug in the cloud too
  - Whenever running at scale in the cloud, you must reduce down to the desktop to understand the results
- Azure is a rapidly moving target and unlike the Grid
  - Commercial cloud backed by large commercial development team
  - Bake in the faults for scaling and resilience

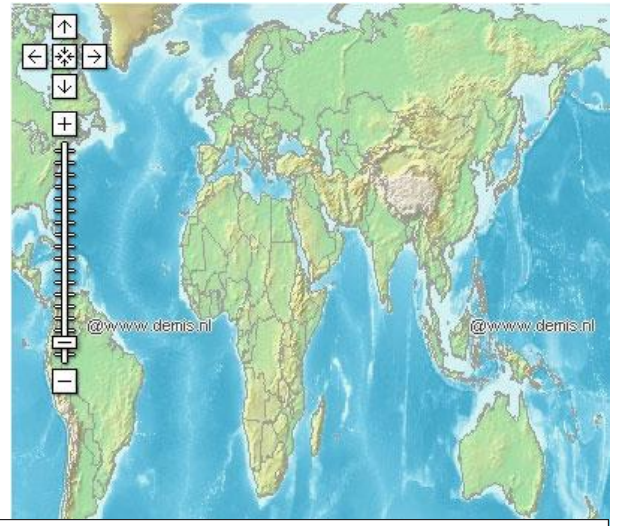


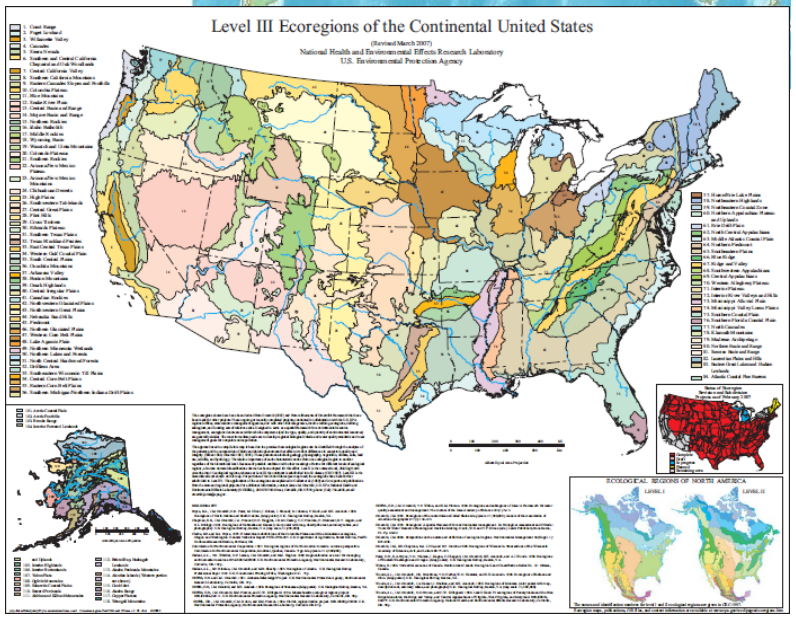


http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx

Space : The Final Frontier

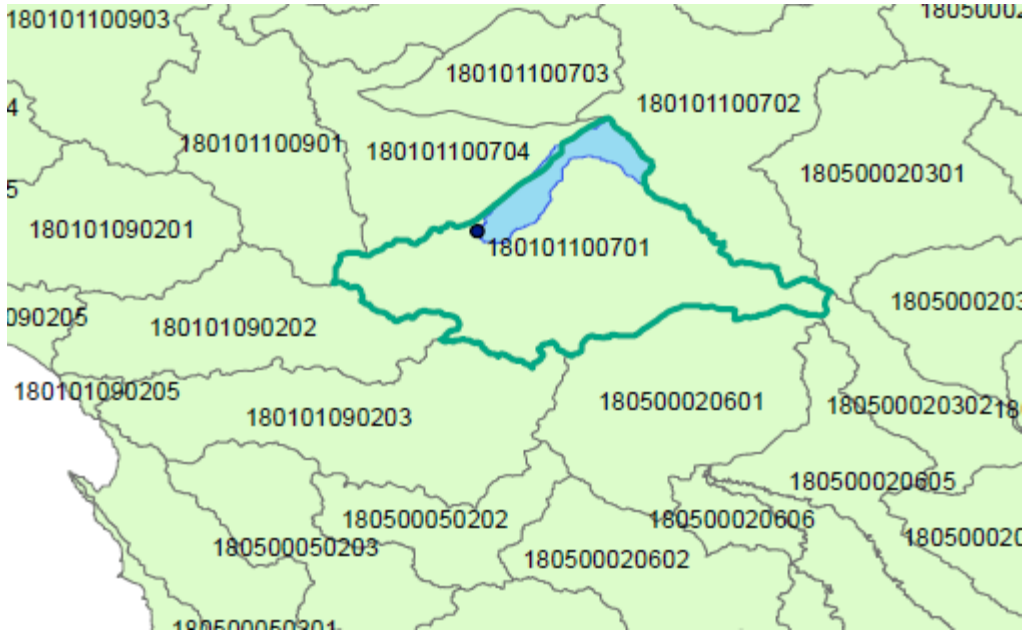
- You are in a maze of twisty little reprojections, all different
  - Swiss army knives exist, but still require download, integration, etc etc.
  - Often must be done with end science in mind
- Categorical spatial classifications driven by both science and human management concerns.
  - Interpretation of data often benefit from cross-discipline mapping





Water Boundary Dataset

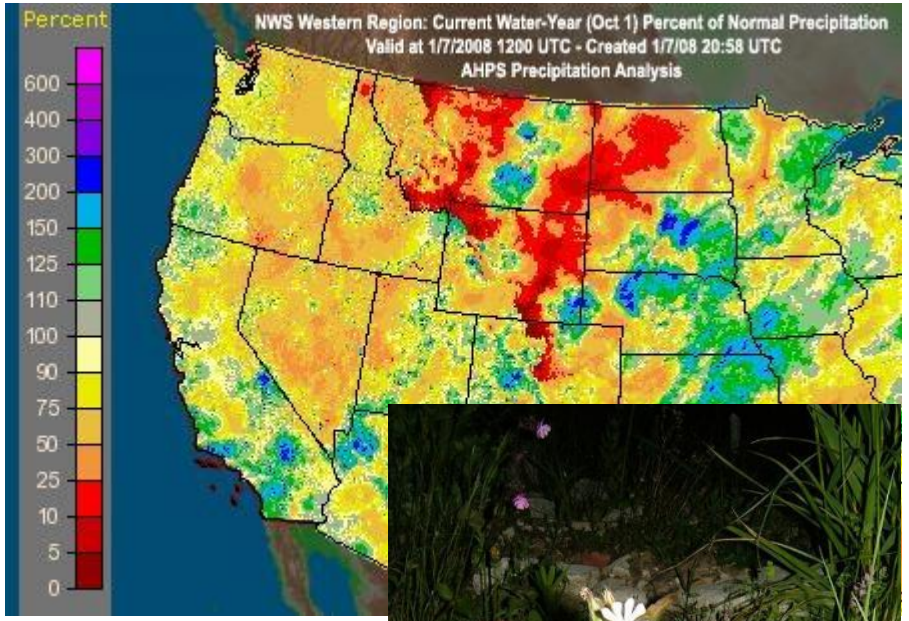
- Hierarchical definition of drainage boundaries across the US
  - 12 year project across USGS
  - 160K sub watersheds
  - Careful and well thought out guidelines for definition across a varying landscape – karsts, mountains, coastal basins, the Great Basin
- Yet not aligned with stream gage locations
  - So computing the water balance across such a watershed is problematic



[Federal Guidelines, Requirements, and Procedures for the National Watershed Boundary Dataset]

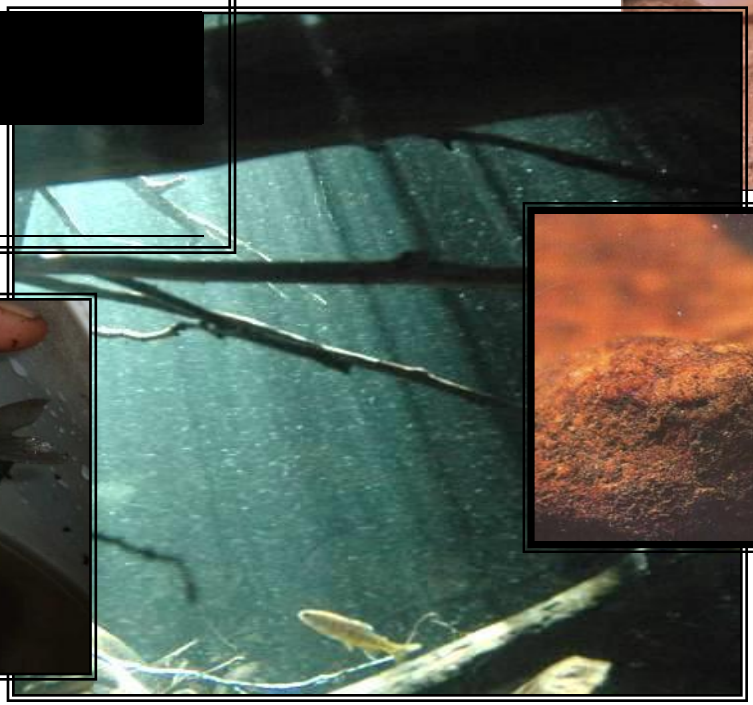
Time is Not Just Another Axis

- Water Year
  - Begins and ends in a relatively dry period
  - In Mediterranean climates, begins 1 October
  - USGS standardizes on 1 October for management reasons
- PAR day
  - Photosynthesis occurs during the day – sensor measurement threshold
  - Solar elevation a close approximation
  - Avoids daylight savings time, time zones



Salmon Month - Different Conditions Different Science

LIFE STAGE	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Adult migration												
Spawning												
Egg Incubation												
Emergence/ Fry												
Juvenile rearing												
Emigration												



Acknowledgements

Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory

- Jim Hunt
- Dennis Baldocchi
- Deb Agarwal
- Monte Goode
- Keith Jackson
- Rebecca Leonardson (student)
- Carolyn Remick
- Susan Hubbard

University of Virginia

- Marty Humphrey
- Norm Beekwilder
- Jie Li (student)

San Diego Supercomputing Center

- Ilya Zavlavsky
- David Valentine
- Matt Rodriguez (student)
- Tom Whitenack

CUAHSI

- David Maidment
- David Tarboton
- Rick Hooper
- Jon Goodman

RENCI

- John McGee
- Oleg Kapeljushnik (student)

Fluxnet Collaboration

- Dennis Baldocchi
- Rodrigo Vargas (postdoc)
- Youngryel Ryu (student)
- Dario Papale (CarboEurope)
- Markus Reichstein (CarboEurope)
- Hank Margolis (Fluxnet-Canada)
- Alan Barr (Fluxnet-Canada)
- Bob Cook
- Susan Holladay
- Dorothea Frank

Ameriflux Collaboration

- Beverly Law
- Tara Hudiburg (student)
- Gretchen Miller (student)
- Andrea Scheutz (student)
- Christoph Thomas
- Hongyan Luo (postdoc)
- Lucie Ploude (student)
- Andrew Richardson
- Mattias Falk
- Tom Boden

North American Carbon Program

- Kevin Schaefer
- Peter Thornton

University of Queensland

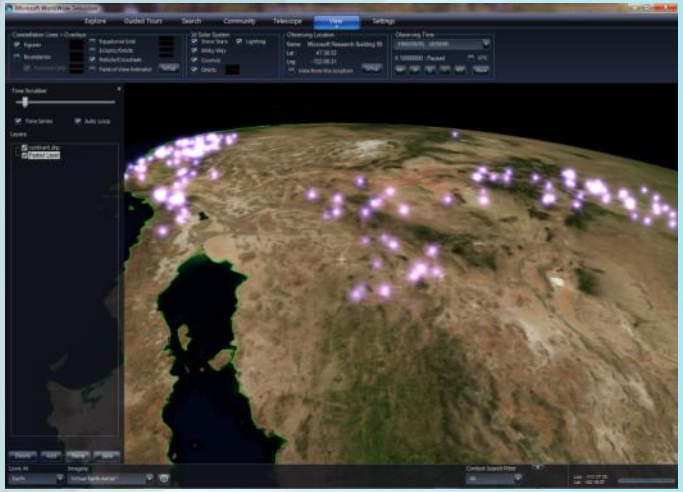
- Jane Hunter

Indiana University

- You-Wei Cheah (student)

Microsoft Research

- Yogesh Simmhan
- Roger Barga
- Dennis Gannon
- Jared Jackson
- Nelson Araujo
- Wei Liu
- Tony Hey
- Dan Fay



Lilac blooms on WWT-E

(Mark Schwartz)

ftp://ftp.ncdc.noaa.gov/pub/data/paleo/phenology/north_america_lilac.txt